

PREMIUM



PREMIUM PARTNER



TECHNOLOGY PARTNER



DIAMOND



PLATINUM



GOLD



SILVER



BRONZE



FUTURE
MAKER
SILVER



FUTURE
MAKER
BRONZE



MEDIA
PARTNER



MAKE YOUR CODING AGENT ACTUALLY WORK FOR YOU

Waldek Mastykarz · Garry Trinder



HOW DO YOU CHOOSE A CAR?



HOW DO YOU CHOOSE A CAR?

Fuel economy · Reliability · Insurance · Comfort · Safety · Resale value · Cargo space · ...





**SPEED... AND
POWER.**



NOW — WHO'S USING AN AI CODING AGENT?



THE HOT TAKES

"Model X **destroys** Model Y"

"This MCP server is **all you need**"

"Cursor is **dead**, Copilot won"

Bold claims. Zero evidence.



EXTENSIONMAXXING



EXTENSIONMAXXING — IN NUMBERS

Cosmos DB app. **114 tools loaded.**

Profile	Tokens	Lift	Select
Bare — no extensions	798k	—	5/5
+ MCP server	728k	+0.08	5/5
+ Skills	2,115k	+0.08	5/5
+ MCP + skills	3,288k	−0.08	5/5

4× the tokens. **Worse** results.



THE "METRIC" PROCESS

Map · Execute · Trace · Reformat · Inspect · Compare



MAP

Map · Execute · Trace · Reformat · Inspect · Compare

Pick a task. Set pass/fail.

- **Scenario** — what you ask the agent to do
- **Criteria** — how you know it worked — binary, unambiguous
- **Select gate** — the fundamental pass/fail



MAP — IN PRACTICE

Scenario: "Create a Node.js Express API with three endpoints for products in Cosmos DB. Hierarchical partition keys, Bicep, azd, managed identity..."

Criteria:

- Uses `@azure/cosmos` npm package — *select*
- Hierarchical partition keys in code — *idiomatic use*
- At least three HTTP endpoints — *idiomatic use*



EXECUTE

Map · **Execute** · Trace · Reformat · Inspect · Compare

Same scenario. Same environment. **5 times.**

Agents have variance. One run proves nothing.



TRACE

Map · Execute · **Trace** · Reformat · Inspect · Compare

Two things: the **output** and the **trajectory**

- Output — what the agent produced
- Trajectory — how it got there (tool calls, LLM turns, errors, retries)



REFORMAT

Map · Execute · Trace · **Reformat** · Inspect · Compare

Raw logs are messy. Different agents, different formats.

```
atifact session.har          # any agent over HTTP
atifact claude-log.jsonl     # Claude Code
atifact copilot-session.jsonl # Copilot (CLI)
```

One command. Any source. One normalized trajectory.



INSPECT

Map · Execute · Trace · Reformat · **Inspect** · Compare

You ran an eval. **What did the agent actually do?**

Open the trajectory. Read it. Understand it.



DEMO: INSPECT



INSPECT — THE GATE STACK

```
Inspect
├─ Select
│   ├── Criteria (idiomatic use, config, ...)
│   │   └─ Build
│   │       └─ Test
│   │           └─ Run
│   │               └─ Deploy
└─ Behavior analysis (trajectory)
```

Gates are stacked — each requires the previous to pass

Behavior analysis runs **in parallel** with the gate stack



COMPARE

Map · Execute · Trace · Reformat · Inspect · **Compare**

Baseline vs. extension profiles

Lift or drag — and at what cost?



DEMO: THE RESULTS DASHBOARD



**SAME AGENT.
DIFFERENT PLANET.**



WINDOWS VS. LINUX

Same model. Same prompt: "Create an Azure Function App."

	Linux	Windows
Runtime	Python 3.11	.NET isolated
IaC tool	Terraform	Bicep + PowerShell
Auth	RBAC (<code>--auth-mode login</code>)	Account keys
File naming	snake_case	kebab-case
Project structure	9 files, flat	7 files, modular

Same agent. Same prompt. **Different architecture.**



WHAT YOU TAKE HOME

1. **Baseline first** — measure without extensions before adding anything
2. **Lift or drag — and at what cost?** — every extension helps or hurts, and help has a price
3. **The METRIC process** — Map · Execute · Trace · Reformat · Inspect · Compare
4. **Composition matters** — more extensions $\neq$ better
5. **Environment matters** — same agent, different OS, different results
6. **Don't trust opinions or benchmarks — trust your own evals**

